1

SYSTEM FOR SYNCHRONIZATION BETWEEN MOVING PICTURE AND A TEXT-TO-SPEECH CONVERTER

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to a system for synchronization between moving picture and a text-to-speech(TTS) converter, and more particulary to a system for synchronization between moving picture and a text-to-speech converter which can be realized a synchronization between moving picture and synthesized speech by using the moving time of lip and duration of speech information.

2. Description of the Related Art

In general, a speech synthesizer provides a user with various types of information in an audible form. For this purpose, the speech synthesizer should provide a high quality speech synthesis service from the input texts given to a user. In addition, in order for the speech synthesizer to be operatively coupled to a database constructed in a multi-media environment, or various media provided by a counterpart involved in a conversation, the speech synthesizer can generate a synthesized speech so as to be synchronized with these media. In particular, the synchronization between moving picture and the TTS is essentially required to provide a user with a high quality service.

FIG. 1 shows a block diagram of a conventional text-to-speech converter which generally consists of three steps in generating a synthesized speech from the input text.

At step 1, a language processing unit 1 converts an input text to a phoneme string, estimates prosodic information, and symbolizes it. The symbol of the prosodic information is estimated from the phrase boundary, clause boundary, accent position, sentence patterns, etc. by analyzing a syntactic structure. At step 2, a prosody processing unit 2 calculates the values for prosody control parameters from the symbolized prosodic information by using rules and tables. The prosody control parameters include phoneme duration and pause interval information. Finally, a signal processing unit 3 generates a synthesized speech by using a synthesis unit DB 4 and the prosody control parameters. That is, the conventional synthesizer should estimate prosodic information related to naturalness and speaking rate only from an input text in the language processing unit 1 and the prosody processing unit 2.

Presently, a lot of researches on the TTS have been conducted through the world for application to mother languages, and some countries have already started a commercial service. However, the conventional synthesizer is aimed at its use in synthesizing a speech from an input text, and thus there is no research activity on a synthesizing method which can be used in connection with multi-media. In addition, when dubbing is performed on moving picture or animation by using the conventional TTS method, information required to implement the synchronization of media with a synthesized speech cannot be estimated from the text only. Thus, it is not possible to generate a synthesized speech, which is smoothly and operatively coupled to moving pictures, from only text information.

If the synchronization between moving picture and a synthesized speech is assumed to be a kind of dubbing, there

2

can be three implementation methods. One of these methods includes a method of synchronizing moving picture with a synthesized speech on a sentence basis. This method regulates the time duration of the synthesized speech by using information on the start point and end point of the sentence. This method has an advantage that it is easy to implement and the additional efforts can be minimized. However, the smooth synchronization cannot be achieved with this method. As an alternative, there is a method wherein information on the start and end point, and phoneme symbol for every phoneme are transcribed in the interval of the moving picture related to a speech signal to be used in generating a synthesized speech. Since the synchronization of moving picture with a synthesized speech can be achieved for each phoneme with this method, the accuracy can be enhanced. However, this method has a disadvantage that additional efforts should be exerted to detect and record time duration information for every phoneme in a speech interval of the moving picture.

As another alternative, there is a method wherein synchronization information is recorded based on patterns having the characteristic by which a lip motion can be easily distinguished, such as the start and end points of the speech, the opening and closing of the lip, protrusion of the lip, etc. This method can enhance the efficiency of synchronization while minimizing the additional efforts exerted to make information for synchronization.

SUMMARY OF THE INVENTION

It is therefore an object of the present invention to provide a method of formatting and normalizing continuous lip motions to events in a moving picture besides a text in a text-to-speech converter.

It is another object of the invention to provide a system for synchronization between moving picture and a synthesized speech by defining an interface between event information and the TTS and using it in generating the synthesized speech.

In accordance with one aspect of the present invention, a system for synchronization between moving picture and a text-to-speech converter is provided which comprises distributing means for multi-media input information, transforming it into the respective data structures, and distributing it to each medium; image output means for receiving image information of the multi-media information from said distributing means; language processing means for receiving language texts of the multi-media information from said distributing means, transforming the text into phoneme string, estimating and symbolizing prosodic information; prosody processing means for receiving the processing result from said language processing means, calculating the values of prosodic control parameters; synchronization adjusting means for receiving the processing results from said prosody processing means, adjusting time durations for every phoneme for synchronization with image signals by using synchronization information of the multi-media information from said distributing means, and inserting the adjusted time durations into the results of said prosody processing means; signal processing means for receiving the processing results from said synchronization adjusting means to generate a synthesized speech; and a synthesis unit database block for selecting required unit for synthesis in accordance with a request from said signal processing means, and transferring the required data.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will become more apparent upon a detailed description of the preferred embodiments for carrying out the invention as rendered below. In the description to follow, references will be made to the accompanying drawings, where like reference numerals are used to identify like or similar elements in the various drawings and in which:

FIG. 1 shows a block diagram of a conventional text-to-speech converter;

FIG. 2 shows a block diagram of a synchronization system in accordance with the present invention;

FIG. 3 shows a detailed block diagram to illustrate a method of synchronizing a text-to-speech converter; and

FIG. 4 shows a flow chart to illustrate a method of synchronizing a text-to-speech converter.

DETAILED DESCRIPTION OF THE INVENTION

FIG. 2 shows a block diagram of a synchronization system in accordance with the present invention. In FIG. 2, reference numerals 5, 6, 7, 8 and 9 indicate a multi-data input unit, a central processing unit, a synthesized database, a digital/analog(D/A) converter, and an image output unit, respectively.

Data comprising multi-media such as an image, text, etc. is inputted to the multi-data input unit 5 which outputs the input data to the central processing unit 6. Into the central processing unit 6, the algorithm in accordance with the present invention is embedded. The synthesized database 7, a synthesized DB for use in the synthesis algorithm is stored in a storage device and transmits necessary data to the central processing unit 6. The digital/analog converter 8 converts the synthesized digital data into an analog signal to output it to the exterior. The image output unit 9 displays the input image information on the screen.

Table 1 as shown below illustrates one example of structured multi-media input information to be used in connection with the present invention. The structured information includes a text, moving picture, lip shape, information on positions in the moving picture, and information on the time duration. The lip shape can be transformed into numerical values based on a degree of a down motion of a lower lip, up and down motion at the left edge of an upper lip, up and down motion at the right edge of an upper lip, up and down motion at the left edge of a lower lip, up and down motion at the right edge of a lower lip, up and down motion at the center portion of an upper lip, up and down motion at the center portion of a lower lip, degree of protrusion of an upper lip, degree of protrusion of a lower lip, distance from the center of a lip to the right edge of a lip, and distance from the center of a lip to the left edge of a lip. The lip shape can also be defined in a quantified and normalized pattern in accordance with the position and manner of articulation for each phoneme. The information on positions is defined by the position of a scene in a moving picture, and the time duration is defined by the number of the scenes in which the same lip shape is maintained.

TABLE 1

Example of Synchronization Information

Input Information	Parameter	Parameter Value
text	sentence	
moving picture	scene	
synchronization information	lip shape	degree of a down motion of a lower lip, up and down motion at the left edge of an upper lip, up and down motion at the right edge of an upper lip, up and down motion at the left edge of a lower lip, up and down motion at the right edge of a lower lip, up and down motion at the center portion of an upper lip, up and down motion at the center portion of a lower lip, degree of protrusing of an upper lip, degree of protrusion of a lower lip, distance from the center of a lip to the right edge of a lip, and distance from the center of a lip to the left edge of a lip
	information on position	position of scene in moving picture
	time duration	number of continuous scenes

FIG. 3 shows a detailed block diagram to illustrate a method of synchronizing a text-to-speech converter and FIG. 4 shows a flow chart to illustrate a method of synchronizing a text-to-speech converter. In FIG. 3, reference numerals 10, 11, 12, 13, 14, 15, 16 and 17 indicate a multi-media information input unit, a multi-media distributor, a standardized language processing unit, a prosody processing unit, a synchronization adjusting unit, a signal processing unit, a synthesis unit database, and an image output unit, respectively.

The multi-media information in the multi-media information input unit 10 is structured in a format as shown above in table 1, and comprises a text, moving picture, lip shape, information on positions in the moving picture, and information on time durations. The multi-media distributor 11 receives the multi-media information from the multi-media information input unit 10, and transfers images and texts of the multi-media information to the image output unit 17 and the language processing unit 12, respectively. When the synchronization information is transferred, it is converted into a data structure which can be used in the synchronization adjusting unit 14.

The language processing unit 12 converts the texts received from the multi-media distributor 11 into a phoneme string, and estimates and symbolize prosodic information to transfer it to the prosody processing unit 13. The symbols for the prosodic information are estimated from the phrase boundary, clause boundary, the accent position, and sentence pattern, etc. by using the results of analysis of syntax structures.

The prosody processing unit 13 receives the processing results from the language processing unit 12, and calculates the values of the prosodic control parameters. The prosodic control parameter includes the time duration of phonemes, contour of pitch, contour of energy, position of pause, and length. The calculated results are transferred to the synchronization adjusting unit 15.